Bayesian estimation applied to multiple species

PHYSICAL REVIEW D **75,** 103508 (2007)

Bayesian estimation applied to multiple species

Martin Kunz[*]
Département de Physique Théorique, Université de Genève, 24 quai Ernest Ansermet, CH–1211 Genève 4, Switzerland

Bruce A. Bassett[†]
South African Astronomical Observatory, Observatory, Cape Town, South Africa
and Department of Mathematics and Applied Mathematics, University of Cape Town, Rondebosch, 7700, Cape Town, South Africa

Renée A. Hlozek[‡]
National Astrophysics and Space Science Programme, University of Cape Town, Rondebosch, 7700, Cape Town, South Africa
(Received 10 October 2006; revised manuscript received 18 February 2007; published 18 May 2007)

Observed data are often contaminated by undiscovered interlopers, leading to biased parameter estimation. Here we present BEAMS (Bayesian estimation applied to multiple species) which significantly improves on the standard maximum likelihood approach in the case where the probability for each data point being "pure" is known. We discuss the application of BEAMS to future type-Ia supernovae (SNIa) surveys, such as LSST, which are projected to deliver over a million supernovae light curves without spectra. The multiband light curves for each candidate will provide a probability of being Ia (pure) but the full sample will be significantly contaminated with other types of supernovae and transients. Given a sample of N supernovae with mean probability, $\langle P \rangle$, of being Ia, BEAMS delivers parameter constraints equal to $N\langle P \rangle$ spectroscopically confirmed SNIa. In addition BEAMS can be simultaneously used to tease apart different families of data and to recover properties of the underlying distributions of those families (e.g. the type-Ibc and II distributions). Hence BEAMS provides a unified classification and parameter estimation methodology which may be useful in a diverse range of problems such as photometric redshift estimation or, indeed, any parameter estimation problem where contamination is an issue.

DOI: 10.1103/PhysRevD.75.103508 PACS numbers: 98.80.Es

I. INTRODUCTION

Typically parameter estimation is performed with the assumption that all the data come from a single underlying probability distribution with a unique dependence on the parameters of interest. In reality the data set is invariably contaminated by data from other probability distributions which, left unaccounted for, will bias the resulting best-fit parameters. This is a typical source of systematic error.

In this paper we present BEAMS (Bayesian estimation applied to multiple species), a method that allows for optimal parameter estimation in the face of such contamination when the probability for being from each of the distributions is known. As a by-product our method allows the properties of the contaminating distribution to be recovered.

For example, the next decade will see an explosion of supernova data with particular emphasis on type-Ia supernovae (SNIa) as standard candles. A few hundred supernovae were known by 2005; see [1–7] and references therein. The current generation of SNe surveys will last to around 2008 and include SNLS [8,9], ESSENCE [10,11], SDSS-II [12,13], CSP [14,15], KAIT [16], CfA [17,18], C-T [19], and SN Factory [20] and will yield of order 10^3 good SNIa with spectra. Proposed next-generation supernova surveys include the Dark Energy Survey [21], Pan-STARRS [22], and SKYMAPPER [23] and will deliver of order 6×10^4 SNIa by 2013, the majority of which will *not* have spectra. Beyond this, the projected ALPACA telescope [24] would find an estimated 10^5 SNIa over three years. The exponential data rush will culminate in the LSST [25,26] which is expected to discover around 2×10^5 SNIa per year, yielding a catalog with over 2×10^6 SNIa multicolor light curves over a ten year period. The vast majority of these candidates will not have associated spectra.

Fortunately recent surveys such as HST, SNLS, and SDSS-II [13,27–29], building on earlier work, have convincingly shown that a probability of any object being a SNIa can be derived from multicolor photometric observations of the candidate. This has become a very active area of research with significant recent advances pursuing a primarily Bayesian approach to the problem [30–33] and suggesting that the future high-quality, multi-epoch light curves will provide accurate (i.e. relatively unbiased) probabilities of being each possible type of supernova (or of not being a supernova at all).

However, since a less than 100% probability of being Ia is insufficient for the standard parameter estimation methodology, these probabilities—no matter how accurate they are—are useless and have been relegated to use in selecting targets for spectroscopic follow-up as it has always

[*]Electronic address: Martin.Kunz@physics.unige.ch
[†]Electronic address: bruce@saao.ac.za
[‡]Electronic address: reneeh@nassp.uct.ac.za

been considered imperative to obtain spectra of the candidates to find Ia's, reject interlopers, and to obtain a redshift for the SNIa.

As a result, even with the relatively small number of supernova candidates today it is impossible to obtain spectra for all good potential SNIa candidates. Instead, only the best candidates are followed up. For LSST and similar telescopes, less than 0.1% of likely SNIa candidates will be followed up spectroscopically. Unfortunately a spectrum for a high-z object is typically very costly to obtain, with the required integration time roughly scaling as $(1+z)^{\alpha}$ with α somewhere between 2 and 6, depending on the specific situation. In practice the situation is more complex since key identifying features such as the Si II absorption feature at a rest frame 6150 Å are redshifted out of the optical at $z \sim 0.4$, requiring either infrared observations or higher signal-noise spectra of the remaining part of the spectrum.

Until now the choices available in dealing with such a flood of candidates were limited. Either one could limit oneself to those candidates with spectra, rejecting the vast majority of candidates, or one could imagine using the full data set—including the contaminating data—to perform parameter estimation. However, undertaking this in a naive way—such as simply accepting all candidates which have a probability of being a SNIa greater than some threshold, P_*—will lead to significant biases and errors that will undermine the entire data set.

In contrast, we introduce in this paper a statistically rigorous method for using the candidates without spectroscopic confirmation for parameter estimation. BEAMS offers a fully Bayesian method for appropriately weighting each point based on its probability of belonging to each underlying probability distribution (in the above example, its probability of being a SNIa, SNIbc, type II etc...). We will show that this leads to a parameter estimation method without biases (as long as the method for obtaining the probabilities is sound) and which improves significantly the constraints on (cosmological) parameters.

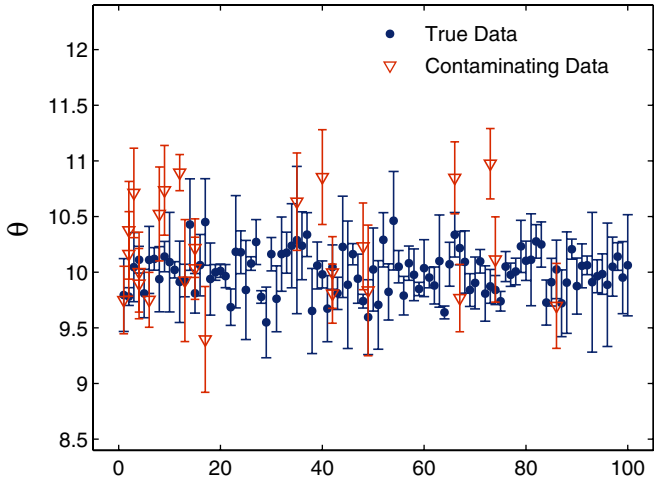

FIG. 1 (color online). Schematic illustration of the problem: data drawn from the true distribution (e.g. type-Ia supernovae) are contaminated by similar looking data from a different distributions (e.g. type-Ibc or II supernovae) leading to biasing in the best fit for parameters and in their errors.

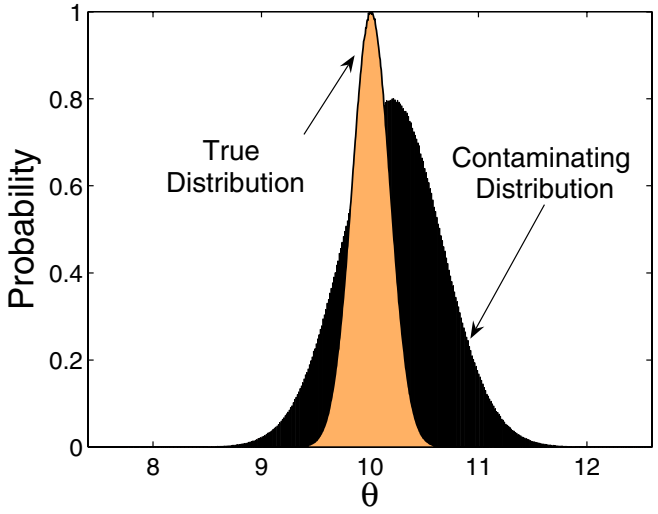

FIG. 2 (color online). Schematic illustration of the underlying distributions used in Fig. 1 for the true and contaminating data. In the case of supernovae, type Ia have a much more narrow intrinsic scatter in their intrinsic luminosity (narrow Gaussian) compared to other types of supernovae (wide Gaussian) which also have different means.

We will be guided by resolving this specific problem, but the underlying principles and methods are more general and can be applied to many other cases. In order not to obscure the general aspects, we will skip over some details, leaving them for future work where actual supernova data are analyzed. We will therefore assume here that we know the redshift of the supernovae (or of its host galaxy), and that we already have estimated the probabilities P_j that the jth supernova is a SNIa (e.g. by fitting the light curves with templates).

To give a simple example, imagine that we wish to estimate a parameter θ (which in cosmology could, for example, represent the luminosity distance to a given redshift) from a single data point, D, which could have come from one of two underlying classes (e.g. supernova type Ia or type II), indexed by $\tau = A, B$ [with their own probability distributions $P(\theta|D, \tau)$, for the parameter θ, as illustrated schematically in Figs. 1 and 2]. Again considering SNe, the link between luminosity and the luminosity distance could be different for the different classes of supernovae due to their intrinsic distribution properties. So given the data D, what is the posterior likelihood for θ assuming that we also know the probability, P_τ, that the data point belongs to each class, τ?

Clearly, $P_A = 1 - P_B$ since we assume the point could come from only one of two classes. Second, as $P_\tau \to 0, 1$, the posterior should reduce to one or other of the class

distributions. Hence, by continuity, the posterior we are seeking should have the form

$$P(\theta|D) = f(P_A)P(\theta|D, \tau = A) + g(P_B)P(\theta|D, \tau = B) \tag{1}$$

where the continuous functions f and g have the limits $f(0) = 0 = g(0)$ and $f(1) = 1 = g(1)$.

Since all the posteriors are normalized we have that $\int P(\theta|D)d\theta = 1 = \int P(\theta|D, \tau)d\theta$. We immediately find that $g(P_A) = 1 - f(1 - P_A)$. The simplest—and as we will show later, Bayesian—choice for f is simply the linear function: $f(P_A) = P_A$. In this case the full posterior simply becomes

$$P(\theta|D) = P_A P(\theta|D, \tau = A) + (1 - P_A)P(\theta|D, \tau = B). \tag{2}$$

This can be easily understood: the final probability distribution for θ is a weighted sum of the two underlying probability distributions (one for each of the classes) depending on the probabilities P_A, $P_B(= 1 - P_A)$ of belonging to each of the two classes.

We will see that our general analysis bears this simple intuition out [see e.g. Eq. (13)].

II. FORMALISM

A. General case

Let us derive in a rather general way the required formulas. Starting from the posterior distribution of the parameters, $P(\theta|D)$, we can work our way towards the known likelihood by repeated application of the sum and product rules of probability theory. The crucial first step involves writing explicitly the marginalization over different data populations, represented by a logical vector τ. Each entry τ_i is either A if the supernova i is of type Ia, or B if it is not. With each entry we associate a probability P_i that $\tau_i = A$, so that the probability for $\tau_i = B$ is $1 - P_i$. For now we assume that these probabilities are known. We can then write

$$P(\theta|D) = \sum_\tau P(\theta, \tau|D) \tag{3}$$

where the sum runs over all possible values of τ. Using Bayes theorem we get

$$P(\theta, \tau|D) = P(D|\theta, \tau)\frac{P(\theta, \tau)}{P(D)}. \tag{4}$$

The "evidence" factor $P(D)$ is independent of both the parameters and τ and is an overall normalization that can be dropped for parameter estimation. We will further assume here that $P(\theta, \tau) \approx P(\theta)P(\tau)$. This simplification assumes that the actual parameters describing our universe are not significantly correlated with the probability of a given supernova to be of type Ia or of some other type. Although it is possible that there is some influence, we can safely neglect it given current data, as our parameters are describing the large-scale evolution of the universe, while the type of supernova should mainly depend on local gastrophysics. In this case $P(\theta)$ is the usual prior parameter probability, while $P(\tau)$ separates into independent factors,

$$P(\tau) = \prod_{\tau_i=A} P_i \prod_{\tau_j=B} (1 - P_j). \tag{5}$$

Here the product over "$\tau_j = A$" should be interpreted as a product over those j for which $\tau_j = A$. In other words, given a population vector τ with entries "A" for SNIa and "B" for other types, the total probability $P(\tau)$ is the product over all entries, with a factor P_j if the jth entry is "A" and $1 - P_j$ otherwise (if the jth entry is "B"). Notice that we discuss here only one given vector τ; the uncertainty is taken care of by the outer sum over all possible such vectors. The full expression is therefore

$$P(\theta|D) \propto P(\theta)\sum_\tau P(D|\theta, \tau) \prod_{\tau_i=A} P_i \prod_{\tau_j=B} (1 - P_j). \tag{6}$$

The factor $P(D|\theta, \tau)$ here is just the likelihood. In general, we have to evaluate this expression, which is composed of 2^N terms for N supernovae. The exponential scaling with the number of data points means that we cannot, in general, evaluate the full posterior—but it should be sufficient to fix $\tau_i = A$ for data points with $P_i \approx 1$ and $\tau_j = B$ for $P_j \approx 0$, and to sum over the intermediate cases. This should give a sufficiently good approximation of the actual posterior.

B. Uncorrelated data

In the case of uncorrelated kinds of data or measurements, such as is approximately true for supernovae [34], we can apply the huge computational simplification pointed out in [40]. In this case, the likelihood decomposes into a product of independent probabilities,

$$P(D|\theta, \tau) = \prod_{\tau_i=A} P(D_i|\theta, \tau_i = A) \prod_{\tau_j=B} P(D_j|\theta, \tau_j = B). \tag{7}$$

The posterior is now a sum over all possible products indexed by the components τ_i. We can simplify it, and bring it into a form that lends itself more easily to the extensions considered in a later section, by realizing that all binomial combinations can be generated by a product of sums of two terms,

$$\sum_\tau \prod_{\tau_i=A} A_i \prod_{\tau_j=B} B_j = \prod_k (A_k + B_k). \tag{8}$$

In this schematic expression, the A_i correspond to the product of likelihood and prior for a $\tau_i = A$ entry, and the B_j to the same product for a $\tau_j = B$ entry. So instead of a sum over 2^N terms, we now only deal with N products.

How do the A_k and B_k look for our supernova application? Let us assume that we are dealing with two popula-

tions, a population A of SNe Ia and a population B of non-Ias. For the kth supernova, A_k is then the product of the probability P_k of being type Ia with the likelihood $P(D_k|\theta, \tau_k = A)$. But since this likelihood is conditional on the supernova being indeed of type Ia, it is just the normal type-Ia likelihood which we will call $\mathcal{L}_{A,k}$. B_k on the other hand is the probability $1 - P_k$ of not being type Ia times the likelihood of the supernovae that are not Ia, which we will call $\mathcal{L}_{B,k}$.

$\mathcal{L}_{A,i}$ is therefore the probability that the ith data point has the measured magnitude if it is type Ia. It is just the usual likelihood, typically taken as a χ^2 in the magnitudes. With the ith supernova data given as distance modulus μ_i and total combined error σ_i (the intrinsic and measurement errors computed in quadrature), it is simply

$$P(D_i|\theta, \tau_i = A) = \mathcal{L}_{A,i}(\theta) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\chi_i^2/2}, \quad (9)$$

with $\chi_i^2 = (\mu_i - m(\theta))^2/\sigma_i^2$ where $m(\theta)$ is the theoretical distance modulus (at redshift z_i). We emphasize that here the normalization of the likelihood is important—unlike in standard maximum likelihood parameter estimation—as we will be dealing with different distributions and their relative weight depends on the overall normalization. In the case of SNe we can of course go a level deeper, since the μ_i are estimated from a number of light-curve points in multiple filters. We could start directly with those points as our fundamental data. Here we ignore this complication while noting that in an actual application this would be the optimal approach [41]

The likelihood $\mathcal{L}_{B,i}$ of a non-Ia supernovae is harder. In an ideal world we would have some idea of the distribution of those supernovae, so that we can construct it from there (see e.g. [42]). If we do not know anything, we need to be careful to minimize the amount of information that we input. It is tempting to use an infinitely wide flat distribution, but such a distribution is not normalizable. Instead we can assume that the non-Ia points are offset with respect to the ''good'' data and have some dispersion. The natural distribution given the first two moments (the maximum entropy choice) is the normal (Gaussian) distribution. The potentially most elegant approach is to use the data itself to estimate the width and location of this Gaussian. This is simply done by allowing for a free shift b and width Σ and marginalizing over them. Optimally we should choose both parameters independently for each redshift bin, in the case where we have many supernovae per bin. Otherwise it may be best to consider b as a relative shift with respect to the theoretical value, modeling some kind of bias.

We would like to emphasize that our choice of the normal distribution for the non-Ia points is the conservative choice if we want to add a minimal number of new parameters. It does not mean that we assume it to be the correct distribution. In tests with a uniform and a χ^2 type distribution for the non-Ia population, assuming a normal distribution sufficed to reliably remove any bias from the estimation process relying on the Ia data points. If we have a very large number of non-Ia points we could go beyond the normal approximation and try to estimate the distribution function directly, e.g. as a histogram. On the other hand, the more parameters we add, the harder it is to analyze the posterior. Also, if we *knew* the true distribution of the contaminants then we should of course use this information. Going back to the full likelihood, we now write [43]

$$P(D|\theta, \tau) = \sum_{b,\Sigma} P(D, b, \Sigma|\theta, \tau) \quad (10)$$

$$= \sum_{b,\Sigma} P(D|b, \Sigma, \theta, \tau) P(b, \Sigma). \quad (11)$$

The last term is the prior on the non-Ia distribution. In the absence of any information, the conventional (least informative) choice is to consider the two variables as independent, with a constant prior on b and a $1/\Sigma$ prior on the standard deviation. In reality, the sum written here is an integration over the two parameters, and the choice of prior is degenerate with the choice of integration measure. As there are no ambiguities, we will keep using summation symbols throughout, even though they correspond to integrals for continuous parameters.

The type-Ia supernovae are independent of the new parameters. They are only relevant for the non-Ia likelihood, which is now for supernova j

$$P(D_j|\theta, b, \Sigma, \tau_j = B) = \mathcal{L}_{B,j}(\theta, b, \Sigma) = \frac{1}{\sqrt{2\pi}\Sigma} e^{-((\mu_j - m(\theta) - b)^2/2\Sigma^2)} \quad (12)$$

(in an actual application to supernova data we would take Σ to be the intrinsic dispersion of the non-Ia population and add to it the measurement uncertainty in quadrature). The posterior, Eq. (6), is then

$$P(\theta|D) \propto \sum_{b,\Sigma} P(\theta)P(b)P(\Sigma) \prod_{j=1}^{N} \{\mathcal{L}_{A,j}(\theta)P_j + \mathcal{L}_{B,j}(\theta, b, \Sigma) \times (1 - P_j)\}. \quad (13)$$

An easy way to implement the sum over b and Σ is to include them as normal variables in a Markov-chain Monte Carlo method and to marginalize over them at the end. Additionally, their posterior distribution contains information about the distribution of the non-Ia supernovae that can be interesting in its own right.

III. A TEST IMPLEMENTATION

In general, θ could of course be a vector of cosmological parameters, but in this section we consider the simple case of the estimation of a constant, corresponding, for ex-

ample, to the luminosity distance in a single bin for the SN case. Continuing with the SN example for simplicity, the data D_i then corresponds to some m_i, an apparent magnitude for each SN in a bin. We again assume that there are two populations, type A (corresponding to SNIa) and type B (everything else).

We fix a distribution for the type A probabilities P_i; for simplicity we take $f(P_i) \propto P_i$, i.e. a distribution that is linearly increasing so that we are dealing predominantly with objects of type A. We then draw a P_i from this distribution and choose an actual type with that probability. Finally, we add a ''spectroscopic'' sample for which $P_i = 1$; i.e. these are guaranteed to be of type A.

We take the type A population to have a known Gaussian distribution with mean $\mu_A = 0$ and variance $\sigma_A = 0.1$. The unknown distribution of type B is taken to be another Gaussian, with mean $\mu_B = 2$ and variance $\sigma_B = 2$. To all data points, A and B, we assign the error bar of type A, i.e. $\sigma_i = \sigma_A$ (but we fit for the error bar of the population B). We assume that this error has been derived e.g. from the dispersion of the spectroscopic sample and that we do not know the distribution of the sample B [44].

The parameters that are being fitted from the data are then μ_A, μ_B, and σ_B, with σ_A fixed from the spectroscopic sample and P_i fixed for each point from an assumed previous step in the analysis (e.g. P_i obtained from goodness of fit to template light curves). As a side remark, although σ_A is assumed here to be known from the dispersion of the spectroscopic sample, it can also be fitted for jointly with the other parameters, which was done in tests of the method [45]; the assumption of fixed known P_i will be relaxed in later sections. To connect this highly simplified example with cosmology, we shall pretend that we consider here only one redshift bin, and that the same analysis is repeated for each bin. The value of μ_A could then be the distance modulus μ in one bin, and an unbiased estimate in all bins would then constrain cosmological parameters like Ω_m, Ω_Λ, etc. The smaller the errors on μ_A, the better the constraints. The data from population B on the other hand give us no information on the distance modulus; hence we must reduce contamination from population B. The posterior that results (explicitly indicating that we estimate μ_A) is then

$$P(\mu_A|D, \sigma_A) \propto \sum_{\mu_B,\sigma_B} \frac{1}{\sigma_B} \prod_{j=1}^{N} [P_j \mathcal{L}_{A,j}(\mu_A, \sigma_A) + (1 - P_j)\mathcal{L}_{B,j}(\mu_B, \sigma_B)], \quad (14)$$

where the mean μ_B and the variance σ_B of population B have taken over the role of the shift b and variance Σ of the last section.

As the population B is strongly biased with respect to A, the algorithm needs to detect the type correctly to avoid wrong results. Table I shows results from an example run with the above parameters, 10 spectroscopic and 1000 photometric data points, where the spectroscopic points are data generated in a Monte Carlo fashion from normally distributed population A and the photometric data consist of points from both population A and population B with associated probabilities $P_i \leq 1$. In this table and all following tables we add a ''Bias'' column that shows the deviation of the recovered parameters from the input values in units of standard deviations.

TABLE I. Example results for the basic algorithm applied to a sample of 10 spectroscopic and 1000 ''photometric supernovae'' in a bin. The bias column shows the deviation from the true value, in units of the standard deviation. A deviation of about 1σ is expected, while about one in 20 realizations is more than 2σ away for random data with normal distribution. BEAMS also allows us to recover the parameters characterizing the contaminating distribution, μ_B and σ_B.

Parameter	Value	Bias (σ)
μ_A	-0.003 ± 0.004	0.8
μ_B	2.00 ± 0.11	0.0
σ_B	1.90 ± 0.07	1.4

For the spectroscopic sample the errors just scale like $\sigma_A/\sqrt{N}$. Each of the other supernova contributes to the good measurement with probability P_j; i.e. each data point has a weight P_j, or an effective error bar $\sigma_A/\sqrt{P_j}$ on average. Defining the average weight

$$w \equiv \frac{1}{N} \sum_{j=1}^{N} P_j \rightarrow \int dP P f(P) = \langle P \rangle \quad (15)$$

where $f(P)$ is the normalized probability distribution function of the P_j, we find that the error on μ scales as

$$\sigma_\mu = \frac{\sigma_A}{\sqrt{N_s + w N_{\rm ph}}} \quad (16)$$

for N_s spectroscopic measurements ($P_j = 1$) and $N_{\rm ph}$ uncertain (photometric only) measurements with average weight w. As can be seen in Fig. 3, the errors on μ recovered by the Bayesian formalism do indeed follow this formula, although they can be slightly worse if the two populations are more difficult to separate than in this example.

In our example where $f(P_j) \propto P_j$ the weight is $w = 2/3$, so that three photometric supernovae equal two spectroscopic ones. The expected error in μ_A for the example of Table I is therefore $0.1/\sqrt{10 + 2/3 \times 1000} \approx 0.004$, in agreement with the numerical result. If we had used only the 10 spectroscopic data points, the error would have been 0.032 so that the use of all available information improves the result by a factor 8. In the case where $f(P_i) \propto (1 - P_i)$, i.e. we are dealing predominantly with type B data, we have a weight of $1/3$. If it is easier to measure three photometric supernovae compared to one spectroscopic one, it will still be worth the effort in this case. We should point out here that these are the optimal errors achievable

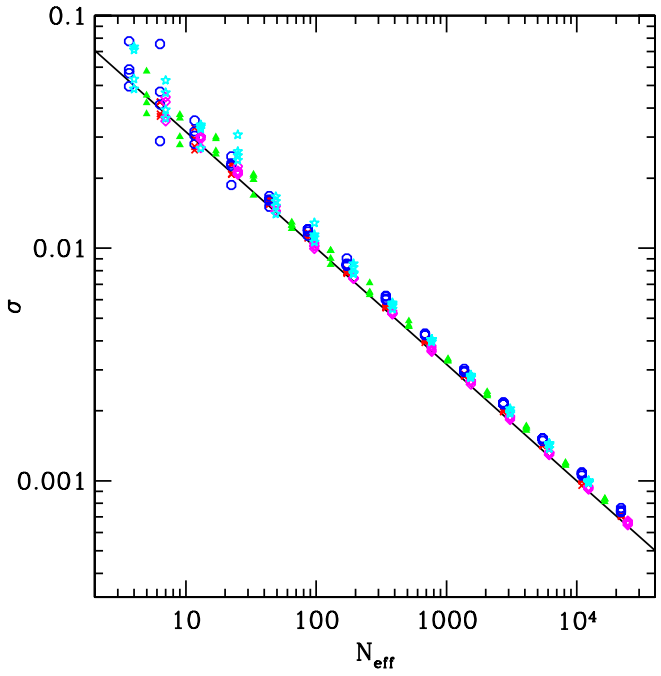

FIG. 3 (color online). Scaling of the errors: The black line shows the expected (optimal) error, which is inversely proportional to the effective number of SNIa given by $\sqrt{N_{\rm eff}}$ in Eq. (17). The different colors and shapes correspond to different distributions of the probabilities P_i (i.e. how many data points have $P_i = 0.9$, how many have $P_i = 0.8$, etc.). The points show the actually measured error for BEAMS given these distributions of the probabilities P_i of the data. BEAMS is able to use nearly all of the information available.

with the data. In Fig. 3 we show the actual recovered error from random implementations with different w and an effective number of SNIa given by

$$N_{\rm eff} \equiv N_s + wN_{\rm ph}. \tag{17}$$

We see that the Bayesian algorithm achieves nearly optimal errors (black line).

We now compare the Bayesian method to some other possible methods:

(i) Use only spectroscopic SNIa.

(ii) Use only SNIa with probabilities above a certain limiting threshold, P. A limit of 0% uses all data points, and a limit of 100% only the spectroscopically confirmed points.

(iii) Weight the χ_i^2 value for the ith point by a function of P_i. This effectively corresponds to increasing the error for data points with lower probability. For the test, we use the weighting $\sigma_j \rightarrow \sigma_j/P_j^{N/2}$. For $N = 0$ this reverts to the limiting case where we just use all of the data in the usual way. For $N > 0$ points are progressively more and more heavily penalized for having low probabilities.

These *ad hoc* prescriptions are not necessarily the only possibilities, but these were the methods we came up with for testing BEAMS against. We now discuss their application to the same test data described above to see how they perform against BEAMS.

Figure 4 shows very clearly that, although the *ad hoc* prescriptions for dealing with the type uncertainty can lead to very precise measurements, they cannot do so without being very biased. Both the Bayesian and the pure-spectroscopic approach recover the correct value (bias less than 1σ), but the latter does so at the expense of throwing away most of the information in the sample.

We can also use BEAMS to get a posterior estimate of the population type, based on the prior value (e.g. from multicolor light curves) and the distribution. To do this for data point j we marginalize over all entries τ *except* τ_j, and additionally over all estimated parameters. In practice this means that we run the program $N + 1$ times, integrating over all parameters (also μ_A in addition to μ_B and σ_B). After an initial run to get the model likelihood for the full solution, we fix in the jth run the jth entry in Eq. (14) to $P_j\mathcal{L}_{A,J}$. Effectively, we compute the model probability if the jth point is assumed to be of type A and compare it to the model probability without this constraint. The relative probability of the two cases then tells us the posterior probability for the model vector τ having the jth entry equal to A, corresponding to the posterior probability of the jth supernova to be of type A. Figure 5 shows an example

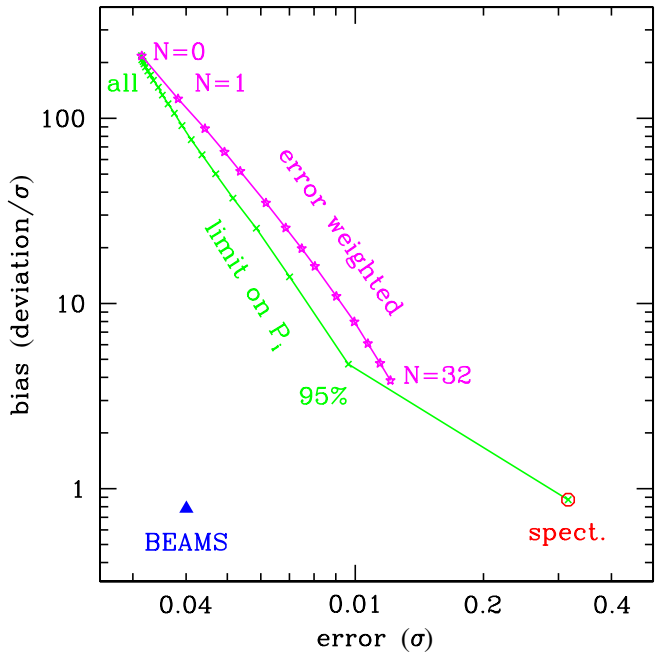

FIG. 4 (color online). A comparison between different methods (see text). Of these methods, only the BEAMS method and the use of the spectroscopic points alone are unbiased. As it can use the uncertain data, the BEAMS method improves the error bars in this example by the (expected) factor of 8. (For the error-weighted method not all values of N were plotted at the high-N end.)

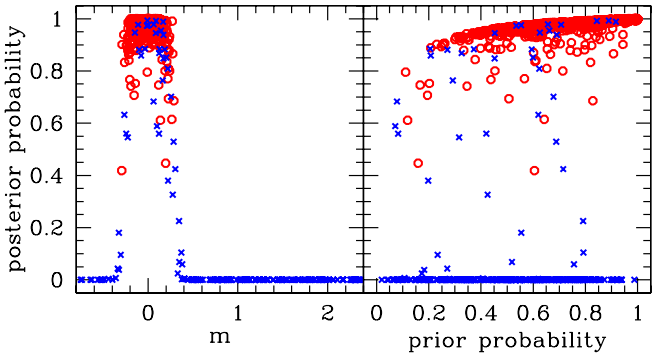

FIG. 5 (color online). BEAMS as a classification algorithm: we plot the posterior probability of the data points to be of type A in our toy example. This depends on their value m_j (left panel) and their prior probability (right panel). The better a point agrees with the recovered distribution of type A, the higher its posterior probability to belong to it. See text for more information. For this test case we know the true nature of the points, and plot population A as red circles and population B as blue crosses (see Table I for distribution characteristics).

case (using a Gaussian approximation to evaluate the integral over all values of the sample mean μ_A). We see how the posterior probability to belong to population A depends both on how well the location of a point agrees with the distribution of A (left panel) and on how high its prior probability was (right panel). In other words, we can reconstruct which points came from which distribution from the agreement between their values of μ and μ_A and their prior probabilities (which is indeed all the information at our disposal in this scenario).

For the toy example the two distributions are quite different, and BEAMS classifies all points within about 3σ of $m = 0$ to be of population A. Here, the prior probability is strongly overwhelmed by the data and the resulting posterior probabilities lie close to 0 and 1 for most data points.

In the following section we extend this basic model in two main directions. First, as reality starts to deviate from the model, there is a danger of introducing a bias. We discuss a few simple cases and try to find ways of hardening the analysis against the most common problems. Second, we extend the model to more than two families, and we also discuss the possibility of using the information on the other populations in the analysis itself.

IV. EXTENSIONS

A. Uncertain probabilities

While the likelihoods used in the estimation of μ_A (which will yield θ) are the same for the earlier example, in this section our treatment and use of the probabilities P_i differ as we begin to include possible error in the P_i's.

Often one may not know the population probability P_j precisely, but has instead a probability distribution. For example P_j may be roughly known, but has an error associated with it (in the SN case this could be due to some systematics in the light-curve fitting process). In this case we have to marginalize over all those probability distributions. For N supernovae this then requires an N-dimensional integration. It is straightforward to include this in a Markov Chain Monte Carlo (MCMC) approach by allowing all P_j to be free variables, but with N of the order of several thousand it may be difficult to get a precise result. On the other hand this may still be better than just sampling P_j at a single point if it is not known exactly.

However, if the measurements are independent, then each integral affects only one of the terms in the product over all data points in Eq. (13). Instead of one N-dimensional integration we are dealing with N one-dimensional integrations which are much easier to compute. In general, we have to integrate each term over the probability p_j with a given distribution $\pi(p_j)$. The case of a known probability corresponds to $\pi(p_j) = \delta(p_j - P_j)$. The next simplest example is the case of a totally unknown probability P_j, for which $\pi(p_j) = 1$. In this case the integral to be solved in each term is

$$\int dp_j(\mathcal{L}_{A,j}p_j + \mathcal{L}_{B,j}(1 - p_j)) = \frac{1}{2}(\mathcal{L}_{A,j} + \mathcal{L}_{B,j}), \quad (18)$$

where $\mathcal{L}_{A,j}$ and $\mathcal{L}_{B,j}$ are the likelihood values of the jth data point assuming population A or B, respectively. The effective probability here turns out to be $P_j = 1/2$. The reason is that we estimate this probability independently for each supernova and do not have enough information to estimate it from the data. In the following subsection we replace this approach instead with a global uncertain probability added to the known distributions. This global probability can then be estimated from the data.

For now, assume we have an approximate knowledge of the type probabilities, say, an independent uncertainty on each P_j, δ_j, so that

$$\pi(p_j) \propto e^{-((p_j - P_j)^2/2\delta_j^2)}, \quad (19)$$

where the proportionality constant is chosen so that the integral over $\pi(p_j)$ from zero to one is 1. If the random error on P_j is small enough that the distribution function is well contained within the domain of integration, i.e. $P_j + \delta_j \ll 1$ and $P_j - \delta_j \gg 0$, then we recover just $p_j = P_j$. In this case the Gaussian distribution function acts effectively as a delta function. For large uncertainties, or for probabilities close to the boundaries, corrections will become important and can bias the result. For the specific case of random errors, the correction term is of the form $\mathcal{L}_{A,j} - \mathcal{L}_{B,j}$. If we suspect large *random* errors it may be worth adding this term with a global prefactor of its own to the full posterior. On the other hand, in real applications we expect that the probabilities close to $P_j = 1$ are quite well known, so that the boundary error is hopefully not too important.

A fixed, common shift is much more worrisome and can bias the results significantly. This can be seen in Table II where we added a systematic shift to the probabilities (enforcing $0 \leq P_j \leq 1$). This is an especially important point for photometric supernova analyses, where dust reddening can bias the classification algorithm. If we do not take into account this possibility, then the analysis algorithm fails because it starts to wrongly classify the supernovae, but hopefully such a large bias is unrealistic.

At any rate, a bias is readily dealt with by including a free (global) shift s into the probability factors of Eq. (13) and by marginalizing over it, resulting in

$$P(\mu_A, \mu_B, \sigma_B|D) \propto \sum_s \frac{1}{\sigma_B} \prod_{j=1}^{N} [\mathcal{L}_{A,j}(P_j + s) + \mathcal{L}_{B,j}(1 - P_j - s)]. \quad (20)$$

It may be a good idea to include such a shift and to check its posterior distribution. Given enough data it does not significantly impact the errors, and it adds stability also in the case of large random uncertainties in the P_j. We found that an additive bias with a constant prior was able to correct all biasing models that we looked at, as is shown in Table III. However, the presence of a significant shift would indicate a failure of the experimental setup and should be taken as a warning sign.

A free individual shift is degenerate with the case of random uncertainties above, as it cannot be estimated from the data, and is not very useful in this context.

B. Global uncertainty

Given how critical the accuracy of the type probability P_j is in order to get correct results, it may be preferable, as an additional test, to discard this information completely. This helps to protect against wrongly classified outliers and the unexpected breakdown or biasing of the classification algorithm.

Even if the probability for a supernova to be either of type Ia or of another type is basically unknown, corresponding to a large error on all the P_i, not all is lost. We can instead include a global probability p that supernovae belong to either of the groups, and then marginalize over it. In this way, the data will pick out the most likely value for p *and* which observations belong to which class. In terms of the posterior (13) this amounts to just replacing all P_j with p and to marginalize over it,

$$P(\mu_A, \mu_B, \sigma_B|D) \propto \sum_p P(p) \frac{1}{\sigma_B} \prod_{j=1}^{N} \times \{\mathcal{L}_{A,j} p + \mathcal{L}_{B,j}(1 - p)\}. \quad (21)$$

The prior on p, $P(p)$, contains any knowledge that we have on the probability that any given supernova in our survey is of type Ia. If we do not know anything then a constant prior works well. As this is a global probability (i.e. all supernovae have the same p), we cannot in this form include any "per supernova" knowledge on p, gained, for example, from spectra or light curves. For this we need to revert to the individual probabilities discussed previously. However, it is a good idea to include the spectroscopic (known to be good) points with an explicit $p = 1$ as they then define which population is the good population and generally make the algorithm more stable.

In our numerical tests with the toy model described in Sec. III, this approach works very well; see Table IV. However if the two distributions are difficult to separate, with similar average and dispersion, then the algorithm can no longer distinguish between them and concludes that the data is compatible with having been drawn from a single distribution with averaged properties. This normally does not lead to a high bias, since otherwise the data would have been sufficient to tease the populations apart. Nevertheless,

TABLE II. Results with a systematic shift (i.e. bias) in the probabilities P_j. Positive shifts lead to a systematic bias in the results, while negative shifts lead to suboptimal errors. However, the negative shifts will bias instead the inferred properties of population B.

Shift of P_j	μ_A	Bias (σ)
+0.1	0.021 ± 0.004	5.5
+0.2	0.128 ± 0.004	31.8
+0.4	0.408 ± 0.004	96.8
−0.4	0.003 ± 0.005	0.6

TABLE III. Same as Table II, but the model allows for a bias (shift) in the P_j. As most supernovae are population A, with $f(P_j) \propto P_j$, the recovered shift grows slower than the input shift. However, it still removes any bias in the estimation of μ_A.

Shift of P_j	μ_A	Bias (σ)	Recovered shift
+0.0	−0.003 ± 0.004	−0.8	0.002 ± 0.011
+0.1	−0.004 ± 0.004	−1.0	0.073 ± 0.012
+0.2	−0.000 ± 0.004	−0.1	0.158 ± 0.015
+0.4	−0.002 ± 0.004	−0.6	0.286 ± 0.016
−0.4	0.004 ± 0.004	1.0	−0.396 ± 0.013

TABLE IV. Same as Table II, but the model uses an estimated global probability p for all supernovae and does not use the P_j (so in reality all runs above are the same). The expected global probability is $p = N_{\text{eff}}/N \approx 0.66$.

Shift of P_j	μ_A	Bias (σ)	Global probability
+0.0	−0.003 ± 0.004	−0.8	0.66 ± 0.02
+0.1	−0.004 ± 0.004	−0.9	0.68 ± 0.02
+0.2	0.000 ± 0.004	0.0	0.66 ± 0.02
+0.4	−0.003 ± 0.004	−0.7	0.64 ± 0.02
−0.4	0.004 ± 0.004	0.9	0.65 ± 0.02

it seems preferable to use the relative probabilities for the supernova types when the information is available and reliable.

C. Several populations

For an experiment like the SDSS supernova survey, a more conservative approach may be to add an additional population with a very wide error bar that is designed to catch objects that have been wrongly classified as supernovae, or those which got a very high Ia probability by mistake.

Of course there is no reason to limit ourselves to two or three populations, given enough data. If we end up with several thousand supernovae per bin we can try to use the data themselves to understand the different subclasses into which the supernovae can be divided.

The expression (13) can be straightforwardly generalized to M classes A_i of objects (for example supernova types) with their own means μ_i and errors σ_i as well as the probability for data point j to be in class A_i of P_j^i,

$$P(\mu_i, \sigma_i|D) \propto \frac{1}{\prod_{i=1}^{M} \sigma_i} \prod_{j=1}^{N} \left\{ \sum_{i=1}^{M} \mathcal{L}_{i,j}(\mu_i, \sigma_i) P_j^i \right\}. \quad (22)$$

For each data point j the probabilities have to satisfy $\sum_i P_j^i = 1$. Of course there has to be at least one class for which the model is known, i.e. for which we know the connection between μ_i and the (cosmological) parameter vector θ (the "Ia" class in the supernova example), or else it would not be possible to use this posterior for estimating the model parameters θ and we end up with a classification algorithm instead of constraining cosmology.

It is possible that we do not even know how many different populations to expect. In this case we can just keep adding more populations to the analysis. We should then also compute the evidence factor as a function of the number M of populations, $P(D|M)$, by marginalizing the posterior of Eq. (4) over the parameters,

$$P(D|M) = \sum_{\theta,\tau} P(D|\theta, \tau) P(\theta, \tau). \quad (23)$$

This is just the integral over all μ_i and σ_i of the "posterior" that we have used so far, Eq. (22), since we did not normalize it. Once we have computed this factor, then we can compare the relative probabilities of the number of different populations by comparing their evidence factor, since by Bayes theorem (again),

$$P(M|D) = P(D|M) \frac{P(M)}{P(D)}. \quad (24)$$

The relative probability of models with m_1 and m_2 populations is then

$$\frac{P(D|m_1)}{P(D|m_s)} \frac{P(m_1)}{P(m_2)} \quad (25)$$

and usually (in the absence of additional information) the priors are taken to be $P(m_1) = P(m_2)$ so that the evidence ratio gives directly the relative probability.

D. Combined formula

What is the best way to combine the above approaches for future supernova surveys? There is probably no "best way." For the specific example of the SDSS supernova survey the probabilities for the different SN populations are derived from χ^2 fits to light-curve templates [29]. We expect three populations, Ia, Ibc, and II, and objects that are not supernovae at all. We expect that last class to be very inhomogeneous, but we would like to keep the supernovae. From the spectroscopically confirmed supernovae we can learn what the typical goodness of fit of the templates is expected to be and so calibrate them. Supernovae where the χ^2 of all fits is, say, 10 higher than for the typical spectroscopic cases are discarded. For the remainder we set $\pi_i = \exp(-(\chi_0^2 - \chi^2)_i/2)$ where χ_0^2 is the typical value for each population. If $\sum_i \pi_i > 1$ then we set the probabilities to be $P_j = \pi_j / \sum_i \pi_i$, otherwise $P_j = \pi_j$. We also write again the more general θ for the parameters of interest. θ can represent, for example, cosmological parameters, or the luminosity distance to a redshift bin. The connection between θ and the data is specified in the likelihoods $P(D_j|\theta, \ldots)$ which, in general, compare the measured magnitude to the theoretical value, with the theoretical value depending on the θ, in other words $P(D_j|\theta, \mathrm{Ia}) = \mathcal{L}_{\mathrm{Ia},j}(\theta)$, and so on. The full formula is then

$$\begin{aligned} P(\theta|D) \propto \sum_{b_k, \Sigma_k} P(\theta) P(b) P(\Sigma) \prod_{j=1}^{N} \{ & P(D_j|\theta, \mathrm{Ia}) P(\mathrm{Ia})_j \\ & + P(D_j|\theta, b_{\mathrm{Ibc}}, \Sigma_{\mathrm{Ibc}}, \mathrm{Ibc}) P(\mathrm{Ibc})_j \\ & + P(D_j|\theta, b_{\mathrm{II}}, \Sigma_{\mathrm{II}}, \mathrm{II}) P(\mathrm{II})_j \\ & + P(D_j|\theta, b_X, \Sigma_X, X)(1 - P(\mathrm{Ia})_j \\ & - P(\mathrm{Ibc})_j - P(\mathrm{II})_j) \}. \end{aligned} \quad (26)$$

If on the other hand we do not trust the absolute values of the χ^2 then we can either add a bias to safeguard against a systematic shift in the absolute probabilities, or allow for a global P_X that an object is not a supernova. For this we always normalize the supernova probabilities to unity, $P_j = \pi_j / \sum_i \pi_i$, and use the likelihood

$$\begin{aligned} P(\theta|D) \propto \sum_{b_k, \Sigma_k, P_X} & P(\theta) P(b) P(\Sigma) P(P_X) \\ & \times \prod_{j=1}^{N} \{ [P(D_j|\theta, \mathrm{Ia}) P(\mathrm{Ia})_j \\ & + P(D_j|\theta, b_{\mathrm{Ibc}}, \Sigma_{\mathrm{Ibc}}, \mathrm{Ibc}) P(\mathrm{Ibc})_j \\ & + P(D_j|\theta, b_{\mathrm{II}}, \Sigma_{\mathrm{II}}, \mathrm{II}) P(\mathrm{II})_j] (1 - P_X) \\ & + P(D_j|\theta, b_X, \Sigma_X, X) P_X \}. \end{aligned} \quad (27)$$

It is probably a good idea to always run an analysis with additional safeguards like this, and preferably a free global bias in the Ia probability, in parallel to the "real" analysis in case something goes very wrong. The global bias Δ might be added as

$$P(\theta|D) \propto P(\theta) \sum_{b_k,\Sigma_k,\Delta_i,P_X} P(P_X) \prod_{k\{\mathrm{Ibc,II},X\}} P(b_k)P(\Sigma_k) \times \prod_{i=1}^{2} P(\Delta_i) \prod_{j=1}^{N} \{[P(D_j|\theta, \mathrm{Ia})[P(\mathrm{Ia})_j - \Delta_1 - \Delta_2] + P(D_j|\theta, b_{\mathrm{Ibc}}, \Sigma_{\mathrm{Ibc}}, \mathrm{Ibc})[P(\mathrm{Ibc})_j + \Delta_1] + P(D_j|\theta, b_{\mathrm{II}}, \Sigma_{\mathrm{II}}, \mathrm{II})[P(\mathrm{II})_j + \Delta_2]](1 - P_X) + P(D_j|\theta, b_X, \Sigma_X, X)P_X\}. \quad (28)$$

In particular, the bias Δ_2 of the Ia vs II probability is useful to catch problems due to dust reddening which can lead to a confusion between these two classes [46].

While estimating a dozen additional parameters is not really a problem statistically if we have several thousand data points, it can become a rather difficult numerical problem which justifies some work in itself. We are using a Markov-chain Monte Carlo code with several simulated annealing cycles to find the global maximum of the posterior, which seems to work reasonably well but could certainly be improved upon.

We notice that, in addition to a measurement of the model parameters θ from the Ia supernovae, we also get estimates of the distributions of the other populations. In principle, we could feed this information back into the analysis. Even though the prospect of being able to use the full information from all data points is very tempting, we may not win much from doing so. We would expect that the type-Ia supernovae are special in that they have a very small dispersion in the absolute magnitudes. As such, they carry a lot more information than a population with a larger dispersion. In terms of our toy example where $\sigma_A = 0.1$ and $\sigma_B = 2$, we need $(\sigma_B/\sigma_A)^2 = 400$ times more population B points to achieve the same reduction in the error. Unless we are lucky and discover another population with a very small dispersion (or a way to make it so), we expect that the majority of the information will always come from the SNIa.

V. CONCLUSIONS

We present a generalized Bayesian analysis formalism called BEAMS that provides a robust method of parameter estimation from a contaminated data set when an estimate of the probability of contamination is provided. The archetypal example we have in mind is cosmological parameter estimation from type Ia supernovae light curves which will inevitably be contaminated by other types of supernovae. In this case light-curve template analysis provides a probability of being a SNIa versus the other types.

We have shown that BEAMS allows for significantly improved estimation when compared to other estimation methods, which introduce biases and errors to the resulting best-fit parameters.

BEAMS applies to the case where the probability, P_i, of the ith point belonging to each of the underlying distributions is known. Where the data points are independent, repeated marginalization and application of Bayes' theorem yields a posterior probability distribution that consists of a weighted sum of the underlying likelihoods with these probabilities. Although the general, correlated case where the likelihood does not factor into a product of independent contributions is simple to write down, it contains a sum over 2^N terms (for 2 populations and N data points). This exponential scaling makes it unsuitable for application to real data where N is easily of the order of a few thousand. This case will require further work.

We have studied in some detail the simple case of estimating the luminosity distance in a single redshift bin from one population consisting of SNIa candidates and another of non-SNIa candidates. In addition to an optimal estimate of the luminosity distance, by including the free shift b and width Σ of the wide Gaussian distribution as variables in the MCMC estimation method, the BEAMS method also allows one to gain insight into the underlying distributions of the contaminants themselves, which is not possible using standard techniques. Provided that the model for at least one class of data is known, this method can be expanded to more distributions, each with its own shift b_i and width Σ_i.

BEAMS was tested against other methods, such as using only a spectroscopically confirmed data set in a χ^2 analysis, using only data points with probabilities higher than a certain cutoff value, and weighting a χ^2 value by some function of the probability. The Bayesian method performs significantly better than the other methods and provides optimal use of the data available. In the SNe Ia case, the Bayesian framework provides an excellent platform for optimizing future surveys, which is specifically valuable given the high costs involved in the spectroscopic confirmation of photometric SNe candidates.

A Bayesian analysis is optimal if the underlying model is the true model. Unfortunately in reality we rarely know what awaits us, and it is therefore a good idea to add some extra freedom to the analysis, guided by our experience. In this way BEAMS can also be applied when the population probability is not known precisely. In this case a global uncertainty is added to the known probability distributions, which can be estimated from the data. In the case of the SNe Ia, one can include a global probability p that the supernovae belong to either group, and then marginalize over it, allowing the data to not only estimate the most likely value for p but also to separate the data into the two classes. This global approach can protect against outliers when the accuracy of the type probability is not known

precisely. It is one of the strengths of Bayesian approaches that they allow one to add quite general deviations from perfect data, which are then automatically eliminated from the final result, and to compute the posterior probability that such surprises were present.

A robust method of application of BEAMS to data from future supernova surveys is proposed to estimate the properties of the contaminant distributions from the data and to obtain values for the desired parameters. Although we have illustrated and developed the BEAMS algorithm here with explicit references to a cosmological application, it is far more general. It can be easily applied to other fields, from photometric redshifts to other astronomical data analyses and even to other fields like e.g. biology. Since it is Bayesian in nature, it can very easily be tailored to the specific needs of a subject, through simple and straightforward calculations.

ACKNOWLEDGMENTS

We thank Rob Crittenden and Bob Nichol for useful discussions and Joshua Frieman, Alex Kim, and Pilar Ruiz-Lapuente for very useful comments on the manuscript. M. K. acknowledges support from the Swiss NSF; R. H. acknowledges support from NASSP.

[1] A. G. Riess, A. V. Filippenko, and P. Challis *et al.*, Astron. J. **116**, 1009 (1998).
[2] S. Perlmutter, G. Aldering, and G. Goldhaber *et al.*, Astrophys. J. **517**, 565 (1999).
[3] M. Hamuy, M. M. Phillips, and N. B. Suntzeff *et al.*, Astron. J. **112**, 2398 (1996).
[4] M. Hamuy, M. M. Phillips, and N. B. Suntzeff *et al.*, Astron. J. **112**, 2391 (1996).
[5] A. G. Riess, R. P. Kirshner, and B. P. Schmidt *et al.*, Astron. J. **117**, 707 (1999).
[6] J. L. Tonry, B. P. Schmidt, and B. Barris *et al.*, Astrophys. J. **594**, 1 (2003).
[7] A. G. Riess, L.-G. Strolger, and J. Tonry *et al.*, Astrophys. J. **607**, 665 (2004).
[8] M. Sullivan (SNLS Collaboration), American Astronomical Society Meeting Abstracts **208**, 58.03 (2006).
[9] P. Astier, J. Guy, and N. Regnault *et al.*, Astron. Astrophys. **447**, 31 (2006).
[10] J. Sollerman, C. Aguilera, and A. Becker *et al.*, ESA Spec. Publ. **637**, 14.1 (2006).
[11] T. Matheson, S. Blondin, and R. J. Foley *et al.*, Astron. J. **129**, 2352 (2005).
[12] M. Sako, R. Romani, and J. Frieman *et al.*, astro-ph/0504455.
[13] http://sdssdp47.fnal.gov/sdsssn/sdsssn.html.
[14] http://csp1.lco.cl/~cspuser1/CSP.html.
[15] M. Hamuy, G. Folatelli, and N. I. Morrell *et al.*, Publ. Astron. Soc. Pac. **118**, 2 (2006).
[16] http://astro.berkley.edu/~bait/kait.html.
[17] M. Hicken, P. Challis, and R. P. Kirshner *et al.*, in American Astronomical Society Meeting Abstracts (2006), p. 72.04.
[18] S. Jha, R. P. Kirshner, and P. Challis *et al.*, Astron. J. **131**, 527 (2006).
[19] M. Hamuy, J. Maza, and M. M. Phillips *et al.*, Astron. J. **106**, 2392 (1993).
[20] G. Aldering, G. Adam, and P. Antilogus *et al.*, in Survey and Other Telescope Technologies and Discoveries, Proceedings of the SPIE, edited by J. A. Tyson and S. Wolff (2002), Vol. 4836, pp. 61–72.
[21] The Dark Energy Survey Collaboration, astro-ph/0510346.
[22] N. Kaiser (Pan-STARRS Team), Bull. Am. Astron. Soc. **37**, 1409 (2005).
[23] B. P. Schmidt, S. C. Keller, and P. J. Francis *et al.*, Bull. Am. Astron. Soc. **37**, 457 (2005).
[24] P. S. Corasaniti, M. LoVerde, and A. Crotts *et al.*, Mon. Not. R. Astron. Soc. **369**, 798 (2006).
[25] J. A. Tyson, in Survey and Other Telescope Technologies and Discoveries, Proceedings of the SPIE, edited by J. A. Tyson and S. Wolff (2002), Vol. 4836, pp. 10–20.
[26] http://www.lsst.org.
[27] A. G. Riess, L.-G. Strolger, and J. Tonry *et al.*, Astrophys. J. **600**, L163 (2004).
[28] L.-G. Strolger and A. G. Riess, Astron. J. **131**, 1629 (2006).
[29] M. Sullivan, D. A. Howell, and K. Perrett *et al.*, Astron. J. **131**, 960 (2006).
[30] B. D. Johnson and A. P. S. Crotts, Astron. J. **132**, 756 (2006).
[31] N. V. Kuznetsova and B. M. Connolly, astro-ph/0609637.
[32] D. Poznanski, D. Maoz, and A. Gal-Yam, astro-ph/0610129.
[33] B. J. Barris and J. L. Tonry, Astrophys. J. **613**, L21 (2004).
[34] Note that this is an idealization for supernovae since at low redshifts supernovae are correlated due to large-scale bulk velocity fields [35–37]. Further, if the supernovae hosts have redshifts estimated from photometry instead of spectroscopy then correlations between SNe will be induced when the 4000 Å break lies in the same filter and will be exacerbated by host extinction issues; see e.g. [38,39]. Since we wish to present the general formalism here we assume that we know the redshifts of the SNe perfectly. In general, the estimation of the redshift must be included in the parameters to be estimated from the data. In addition to these challenges, the template correction is usually computed using the supernova sample itself and may introduce some correlations. But, in general, if the property of supernovae which makes them standard candles depends *on the sample* then we should be worried. So here we assume that the template corrections and errors were derived previously with the spectroscopic supernovae.

Indeed, strictly speaking, we should use a *different* sample for that purpose, or else estimate those parameters as well as the global dispersion *simultaneously* with the cosmological parameters. In the latter case it is important to keep the $1/\sigma$ normalization of the likelihood and to use additionally a ''Jeffreys prior'' $\propto 1/\sigma$ to avoid a bias towards larger dispersions. In the case where correlations are important one must compute the full probability which is computationally intense, though systematic perturbation theory may be useful for small correlations.

[35] A. Cooray and R. R. Caldwell, Phys. Rev. D **73**, 103002 (2006).

[36] C. Bonvin, R. Durrer, and M. Kunz, Phys. Rev. Lett. **96**, 191302 (2006).

[37] L. Hui and P. B. Greene, Phys. Rev. D **73**, 123526 (2006).

[38] A. G. Kim and R. Miquel, Astropart. Phys. **24**, 451 (2006).

[39] D. Huterer, A. Kim, and L. M. Krauss *et al.*, Astrophys. J. **615**, 595 (2004).

[40] W. H. Press, astro-ph/9604126.

[41] We thank Alex Kim for pointing this out to us.

[42] D. Richardson, D. Branch, and D. Casebeer *et al.*, Astron. J. **123**, 745 (2002).

[43] Again we stress that for the sake of clarity and generality we assume that we know the redshifts of the SN perfectly. If the redshift must also be estimated from the data then the formalism below must be extended in the obvious way.

[44] Although this specific example considers a Gaussian distribution for the m_i of the ''non-Ia'' population B, which corresponds to its likelihood, we have also tested the algorithm for other distributions.

[45] In this case its prior needs to be $\propto 1/\sigma_A$ to avoid biases.

[46] D. Poznanski, A. Gal-Yam, and D. Maoz *et al.*, Publ. Astron. Soc. Pac. **114**, 833 (2002).